研究論文

Student Ratings of Faculty and Courses : Investigating Instructor Effectiveness

Christine Winskowski

INTRODUCTION

"The unexamined life is not worth living." (Socrates)

"Poor evaluation, whether of students or of faculty, renders an unfair judgment and fails to reveal shortcomings in performance. Good evaluation, on the other hand, provides decision makers with the information necessary for informed choices, and teachers with useful feedback for improvement." (Centra, 1993, p.1)

Student evaluation of college and university courses and faculty has recently been introduced in Japan. This has been part of a larger picture of change in tertiary education, change which has included the shift from government-supported institutions to privatized institutions. This shift has in turn has prompted the call for greater accountability in the educational processes, among other things. A focus on teaching effectiveness at the individual and institutional levels is one result. Student ratings forms, administered at the end of a course, are intended to provide a key measure of faculty effectiveness in delivering a course.

These rating or evaluation forms, typically administered on one of the last days of class, before the final examination, have long been part of university and college education in the United States, Australia, Canada, Britain, Europe, and Asia (Haskell, 1997a; Wilson, 1998). When I was a university student in the 1970's, end-of-course evaluations were routinely administered in every class, and throughout my career, I have taken their presence for granted. In the U.S., student ratings of courses are primarily used to assess teaching effectiveness, as well as in the promotion and tenure decisions and determination of annual salary increases for faculty members.

Student evaluation of courses and faculty was introduced in a few U. S. institutions in the mid-1920's. However, it was during the social ferment of 1960's that the use of such course evaluations began to be widely accepted. A commonly cited study published by Seldin (1993) claimed that the percentage of colleges and universities using student evaluations of faculty rose from 29% in 1973 to 86% in 1993. The figure is almost certainly higher today, since shrinking student populations and tighter budgets have underscored the need for improving and ensuring a high quality of teaching.

What may be surprising to colleagues in Japan and other countries is the degree of controversy surrounding the use of students' ratings of courses and faculty. There is a large and sometimes controversial literature on this topic. Cashin estimated in 1995 that there were more than 1500 articles and books on the subject (other estimates are as high as 2000; see Felder, 1992: Wilson, 1998). What are called "student ratings of instructors" (SRI), "student evaluations of faculty" (SEF), etc., are regarded as the best researched of all personnel instruments. However, not all agree that student ratings have uniformly positive effects.

One purpose of this article is to introduce generally what is known about the use of student ratings forms from the literature in the U.S. to colleagues in Japan, including dimensionality, reliability and validity, variables affecting ratings, controversies, and best (recommended) practices.

There is a second purpose as well. Because student ratings of faculty and courses has undergone such extensive research and development, ratings forms have become quite standardized. These standard student ratings forms offer the benefits of reliability, validity, and an extensively researched framework of known variables. The information offered by such forms is necessarily vague and general. That is, if an "overall course" rating is low, it tells us there is need for

improvement; it does not tell us how to improve, or where to look for a problem. For faculty seriously interested in investigating their own teaching-learning processes in the classroom, I propose an alternative approach of individually designed ratings forms for specific class purposes. Faculty who design their own ratings forms can find out what they really need and want to know about events in the classroom. Such ratings forms may complement use of standard ratings forms.

The educational contexts of Japan and the United States are, of course, different. Differences of history, society, and educational philosophy may ensure that some issues and concerns on the topic of student rating of faculty and courses pertinent in the U.S. have little relevance to Japan. At the same time, given the substantial history and research on ratings instruments and their impacts on the educational process, some knowledge of the issues, benefits, and pitfalls of conventional student ratings may be of interest. Research on the variables that are known to influence student ratings in the U.S., for example, may help us to identify potential lines of research in Japan. Examination of best practices may assist determining whether such practices are applicable for Japan. Controversy on validity and fairness issues may illuminate the value of student evaluation to faculty members and administrators. As classroom and teaching processes come under increasing scrutiny, it is the faculty member who is the key to controlling teaching effectiveness.

The ultimate purpose of this piece is to help faculty members understand both the power and limitations students' evaluation of teaching for the development of teaching effectiveness, and to wield student ratings instruments with as much effectiveness as their other teaching tools.

Uses of Ratings Forms

It is widely agreed that the original use for student ratings of faculty and courses was to improve the effectiveness of a faculty member's teaching. This is referred to as a *formative* purpose in the student ratings literature. In the U.S., in recent years, a second major use is (together with other documents) the evaluation of a faculty member for promotion and tenure, and for salary increases. This latter use is referred to as a *summative* purpose. Centra (1993) advises that it is probably best for an institution to use student evaluations first for formative purposes, and then for summative purposes. This allows faculty to become accustomed to the processes and procedures, and to work on teaching improvement before evaluations of teaching are used for administrative decision-making. Issues on student ratings sometimes relate to whether they are being considered for formative or summative purposes.

Dimensions of Rating Forms

What are student ratings forms made of? And what do they measure? It must be expected, of course, that student ratings instruments must have *face validity*, that is, the items used must appear to evaluate effective teaching. But does effective teaching refer to one general dimension of teaching behavior? Or is it made of several sub-dimensions? And do ratings instruments measure one dimension? Or several?

Some universities and colleges in the U.S. use well-established and well-researched student ratings forms such as the Student Evaluation of Educational Quality (SEEQ) (Marsh, 1976), the Student Instructional Report II (SIR II) from Educational Testing Service (ETS, 2004), or the Instructional Development and Effectiveness Assessment (IDEA., no date) of Kansas State University's Center for Faculty Evaluation and Development. These instruments and others like them have both general teaching effectiveness items and items on specific teaching behaviors. McKeachie (1997) notes that for purposes of instructor improvement, general ratings of course satisfaction or instructor effectiveness are not very helpful. Specific items are more likely to give useful feedback to an instructor. Thus ratings forms are likely to measure a number of dimensions of teaching.

Cashin (1995), citing Braskamp and Ory (1994), and Centra (1993) observed that there were six dimensions of teaching behavior frequently found in large studies. They are:

1. Course organization and planning
2. Clarity, communication skills [of instructor]
3. Teacher-student interaction and rapport
4. Course difficulty and workload
5. Grading and examinations

6. Students' rating of their own learning

The SEEQ (Marsh, 1976) has these dimensions:

1. Students' learning
2. Instructor's enthusiasm
3. Group interaction [how students are encouraged to participate]
4. Course organization
5. Instructor's rapport with students
6. Breadth of coverage [of the subject]
7. Value of assignments
8. Examinations
9. Overall course and instructor ratings
10. Course difficulty, workload, pace and other characteristics

The SIR II covers these dimensions, according to its introductory pages (ETS, 2004):

1. Supplementary instruction methods
2. Student effort and involvement
3. Communication [presumably the instructor's]
4. Assignments, exams, and grading
5. Overall evaluation
6. Student comments
7. Course outcomes
8. Course organization and planning
9. Faculty/student interaction
10. Course difficulty, workload, and planning
11. Student information
12. Supplementary questions

Readers can see that many of these dimensions found in one rating form are also found in a second or third form. According to d'Apollonia and Abrami (1997), these are dimensions that are "important to teaching" and "can be judged by students" (p. 1198).

However, these same authors point out that in the past 30 years, teaching styles [in the U.S.] have become more diverse. Moving from the traditional lecture format, faculty have now begun to use interactive seminars, laboratories, cooperative learning, computer-assisted instruction, individualized instruction, and internships (d'Apollonia and Abrami, 1997, p. 1199). Undoubtedly for this reason, some institutions have moved from using a single "omnibus" (i.e. one type fits all) rating form to using multiple versions of a student rating form. One example is Kansas State University's IDEA, which is tailored to the faculty member's teaching objectives (IDEA Center, n.d.). Another example is University of Tennessee, which has developed 11 versions of their Student Assessment of Instruction System (SAIS, 2003), each designed for a different type of class. The individual SAIS forms are designed for small lecture/discussions, large lectures, seminar discussions, problem-solving classes, clinical or other experiential learning, teaching assistant-led discussion sections for a large lecture, lab science classes, among others (SAIS, 2003). These forms vary three to six of the 22 questions about the course, using items that are directed to the specific type of the course.

Groups of items in a student rating form, as can be seen above, are sometimes labeled to reflect a specific dimension of teaching behavior. Researchers have used factor analysis in an attempt to discover if such groups of items actually do measure specific, independent dimensions of teaching behavior observable by students.

Factor analytic studies. Factor analytic studies have been used with student rating forms to see if they measure a single global "instructor effectiveness" dimension, or different dimensions of teaching reflected in the items; that is, when one item increases (or decreases), others predictably do so too. If so, it would suggests that the forms reflect specific components or dimensions of teaching effectiveness. In some cases, these studies may be intended to explore whether a rating instrument with ungrouped and unlabeled items actually have a subset of items, that vary together, called a factor. In other cases, these studies try to confirm that a rating form like the SEEQ has the dimensions that it claims to. In still other cases, several ratings forms have been investigated to see whether there is some consistent global dimension evident across many forms.

A review of these studies by d'Apollonia and Abrami (1997) reveals several things: The authors note, for example, that several studies (e.g. by Herbert Marsh) have indeed confirmed the factor structure of the SEEQ form. However, from their own and others' prior factor analytic investigations, the authors argue that while factors point to item clusters which co-vary, this is not absolute evidence that the item clusters in fact represent

dimensions of teaching effectiveness that are socially and psychologically real. That is, the apparent factors may be an artifact of the factor analysis. Additionally, they point out that various methods of factor analysis result in various factor solutions showing various item groupings, with no clear indication of which one is the best or "correct" one.

In describing a re-analysis of their own work (published in 1993 and 1996), d'Apollonia and Abrami (1997) indicate that they examined 17 student rating forms, coded the 458 items in these ratings forms into common instructional categories, produced a correlation matrix (where each item was correlated with all other items), and factor analyzed the matrix to see which items varied together. They found a large principle component (factor) across the rating forms which accounted for 63% of the variance in students' responses. Then they found four sub-skills which could be seen as reflecting these dimensions of teaching behavior: 1) The instructor's delivery of information, 2) instructor's facilitation of interaction with students, 3) instructor's evaluation of student learning, and 4) a collection of miscellaneous teacher behaviors. The first three skills were similar to those found by other researchers. This could be taken as confirmation of multiple dimensions in the rating of teacher effectiveness. However, since these sub-skills are interrelated (i.e. not independent from each other statistically) the authors conservatively contend that an equally viable interpretation of the factor studies is that the sub-factors are artifacts of the measurement process, and that "student rating forms measure a global component, General Instructional Skill" (p. 1201). Thus, while many researchers are convinced that there are dimensions of teacher effectiveness, and that they are demonstrably measurable, the matter remains unresolved.

Reliability of Rating Forms

Reliability, or inter-rater reliability, of student ratings generally refers to the degree to which student raters in a class agree with one another on a given item. Strong inter-rater reliability indicates that there is a stability and consistency in the students' responses to the course evaluation items. Reliability in research on student ratings is usually calculated as a correlation coefficient, and the median coefficient across a class of students is taken as the reliability coefficient.

Reliability in student ratings is one of the least controversial findings. Cashin (1995) found median reliabilities of .69 with 10 raters, .83 with 15 raters, .88 with 30 raters, and .91 with 40 student raters or more. Ratings taken at the end of a course, and one year or more later were stable, showing an average correlation of .83 (Overall and Marsh, 1980, cited in Cashin, 1995; Marsh and Dunkin, 1997, cited in Paulsen, 2001).

Validity of Rating Forms

Recall that the primary purpose of student ratings of faculty and courses is as a measure of the faculty member's teaching effectiveness. Validity of student ratings refers to whether a rating instrument in fact measures teaching effectiveness, as it is intended to. This is a large and complex area with several aspects. One recurring theme in many writings is that there is no universal agreement on what constitutes "teaching effectiveness." Validity is also an area that seems to have attracted the lion's share of controversy. However, Greenwald (1997) notes in his overview and introduction to a special issue of *American Psychologist* on student ratings that the majority of publications in the preceding 25 years support the argument that student ratings are valid measures of teaching effectiveness (p. 1182).

Multisection validity studies. A straightforward indication of teaching effectiveness is how much or how well students learn. A straightforward indication of how well students learn is their grades. For this reason, there are many studies on the relationship between student ratings and their grades. A common research design, called a multisection validity study, compares large courses with many sections, each following the same syllabus and using the same examinations. Over 40 studies using the multisection methodology have been identified by d'Apollonia and Abrami (1997). Because it is such a large literature, meta-analytic reviews have also been undertaken to determine the overall trends in the research. d'Apollonia and Abrami point out that "not only do the primary researchers reach opposite conclusions about the validity of student ratings, but so do some of the reviewers" (1997, p. 1201).

One reason for this may be found in a re-analysis of 43 multisection validity studies (Abrami, d'Apollonia, and

Cohen, 1990). The authors found that the 43 studies were quite different in what variables they addressed and controlled, including: Aspects of the rating form itself (how the scales were constructed, reliability of the scales, how the scales were administered), aspects of the student achievement measure (i.e. use of an exam, its length, method of scoring, learning criteria, reliability, and administration), aspects of the instructor (personality, experience, sex, rank), aspects of the students (major, personality, ability), aspects of the course (level, subject, whether required or elective), and many other aspects, all totaling 75. Abrami, d'Apollonia, and Cohen confirmed that the average validity coefficient for these instruments is moderatively positive, however they found that "the result of these studies appear inconsistent both from study to study and across rating dimensions".... Characteristics that predict the relationship between student perceptions of teaching and instructor effects on learning vary with the aspect of teaching being investigated" (1990, p. 230). In a later meta-analysis published in 1996, d'Apollonia and Abrami again conclude that, after statistically correcting for methodological differences, differences in reliability, etc., the correlation between student ratings of "general instructional skill" and student learning was moderate to large, namely .47 (1997, p. 1202). Most other (though not all) researchers and reviewers agree. Kulik (2001), reviewing other meta-analytic studies, cites Cohen's 1981 review of 41 studies on 68 courses with multiple sections. Cohen found the average correlation between examination scores and overall teacher evaluation was also .47. Kulik concluded that this was a moderate to high correlation (2001, p. 12). (See also Cashin, 1995; Felder, 1992; and numerous others, e.g. those cited in Haskell, 1997a.)

The implication seems to be that while most accept that the validity of student ratings of their instructors is moderately positive, there may be a good deal more complexity in the classroom setting and the features which affect and comprise teaching effectiveness than we presently know.

Multitrait-multimethods validity studies. Cashin (1995) identified other, less direct indications of teaching effectiveness that had correlations with student ratings. Cashin lists these studies (some primary research and some reviews of other studies), their criteria of teacher effectiveness, and their correlations with student ratings:

Effectiveness Criterion	Authors	Correlation
self rating of instructor	Feldman, 1989	average .29
	Marsh, Overall, & Kesler, 1979	.49
	Marsh and Dunkin, 1992	median .45
administrator's ratings of instructor	Kulik & McKeachie, 1975	range .47 - .62
	Feldman, 1989	average .39
colleagues' ratings of instructors	Kulik & McKeachie, 1975	range .48 - .69
	Feldman, 1989	average .55
alumni ratings of instructors	Overall & Marsh, 1980; and Braskamp & Ory, 1994	range .40 - .75
	Feldman, 1989	average .69
trained observers's ratings of instructors	Feldman, 1989; and Marsh & Dunkin, 1992	average .50
students' written comments	Ory, Braskamp & Pieper, 1980	.93
	Braskamp, Ory & Pieper, 1981	.75

Table 1. Correlations between student ratings and measures of teacher effectiveness (excluding grades), based on Cashin 1995

This evidence does provide some support for the claim of student ratings' validity, even though these effectiveness criteria are considered less compelling than that in multisection studies. Abrami, d'Apollonia, and Cohen (1990) conclude that "the MTMM [multitrait-multimethod] validation designs provide weaker evidence for the validity of student ratings as measures of instructional effectiveness than do multisection validation designs" (p. 221). It should be noted that researchers may differ in their interpretation of what constitutes strong, moderate, and weak correlations. Also, a correlation of a given value may be more significant for one criterion of teacher effectiveness than another. However, one cannot fail to be impressed with the consistency of these correlations, and their relative size (all but two are .40 or higher).

Between the multisection validation studies and multitrait-multimethod validation studies, we can see why students' rating of instructors is generally accepted by most as a valid reflection of overall teaching effectiveness.

Bias in Student Ratings

What variables could introduce bias into the student rating process, i.e. what could influence the students' ratings, even though the influences are unrelated to teaching effectiveness? This question reflects the serious concern faculty members have over the fairness and validity of student ratings of teaching. In response, there has been a wealth of research done to examine instructor

variables, student variables, course variable, and other variables. Cashin, in a review of student ratings research (1995), summarizes what is known about potential relationships between student ratings and other variables. Ory (2001) provides a similar view. This section is based primarily on their articles, except where otherwise noted.

Variables having little or no relationship to student ratings.

1) There is little or no correlation between student ratings and the age or experience of the instructor. Where small differences were observed, they tended to be negative, i.e. older faculty received slightly lower ratings. However, longitudinal studies following the same faculty for up to 13 years confirmed this lack of relationship.

2) A very low correlation of .12 has been found between instructors' research productivity and student ratings. Cashin concludes that research productivity reflects neither good nor bad teaching.

3) Students' age, gender, year (e.g. freshman, sophomore, etc.), grade-point average, and personality appear to be unrelated to student ratings.

4) Smaller classes tend to received higher student ratings than large ones, but the relationship is quite small (Cashin cites two studies showing negative correlations of -.14 and -.09). It appears that size has little significant effect.

5) Finally, there is no relationship between the time of day that a course is taught and student ratings.

Variables which may (or may not) have a relationship to student ratings.

1) Among studies reported by Cashin (1995) and by Ory (2001) examining the gender of instructors, a very slight differences (r = .02) in favor of women was found in one review of 28 studies. In a few other studies, students of each gender rated same-gender instructors more highly then opposite-gender instructors. However, results may be more complex than they first appear. In a series of studies that Sinclair and Kunda (2000) conducted in university classrooms and in an experimental laboratory, they found that students rated male and female instructors equally highly when they had received high grades or other positive evaluation. However, when students received low grades or negative evaluation, they rated female instructors lower than male instructors. The authors theorize that when people receive negative evaluation, they will employ stereotypes (e.g. that women are not as competent as men) to preserve their view of themselves.

2) There is little research on the relationship of instructor's race and student ratings, but what there is indicates that race is not related to student ratings, according to Cashin (1995). However, some work by Sinclair and Kunda (1999) suggests that in their studies, people praised by a black professional seemed to inhibit negative stereotypes about them; those criticized by the black professional did not.

Both of these findings, if confirmed by further research, could have serious implications for ratings. Fich, in responding to the conflicting conclusions to be found in these studies notes that "the problem is that there are many variables unrelated to the quality of teaching that may affect evaluations and that interact in complex ways. Furthermore, most of this work consists of statistical analyses, where factors that are significant for a small segment of the population, for example, women computer science professors, can be insignificant in the aggregate data" (Fich, 2003, p.2).

Variables which do have relationships to student ratings.

1) Regular faculty tend to get higher student ratings than graduate teaching assistants, as might be expected.

2) There is some evidence that instructor expressiveness in delivering the course content appears to enhance learning, and thus receives higher student ratings. However, this research has not convinced all critics.[1] Still, Cashin points out that making class interesting as well as informative promotes learning (1995, p. 5).

---

[1] For example, many writers have referred to the "Dr. Fox Effect," from a study by Naftulin, Ware, and Donnelly (1973), in which an enthusiastic lecture delivered by "an expert" (actually an actor), received high ratings. Critics of student ratings argue that this study showed that students will rate entertaining lectures highly, even though the lecture was actually full of misinformation. Other researchers point out that the study of a single staged lecture is not representative of a real classroom setting.

3) Some personality traits of the instructor appear to have small but significant correlations with students' ratings: positive self-esteem (r = .30) and energy and enthusiasm (r = .27). One intriguing study of different psychology classes showed that a) teaching effectiveness varies across types of classes (e.g. freshman lectures, graduate seminars); b) student ratings were strongly related to peer ratings of instructor personality traits; c) ratings correlated most strongly with the traits of leadership, extraversion, liberalism, supportingness, intellectual curiosity, and changeableness; and d) that personality traits which contributed to effective teaching "varied substantially for different types of courses" (Murray, Rushton, and Paunonen, 1990, p. 259). [For reasons that are unclear, Cashin has placed these findings in the section on variables showing "little or no relationship to student ratings." The placement of these findings in the section on ratings-related variables is the author's.]

4) Elective courses get higher student ratings than required courses. Student ratings are positively associated with student motivation (average correlation = .40), and general interest (i.e. when courses are selected as electives). Cashin (1995) notes that this is a variable requiring control for research purposes. Ory (2001) comments that on his campus, separate norms have been developed for required, elective, and mixed required-elective courses.

5) Several researchers have found a consistent positive correlation of .10 to .30 between student ratings and expected grades, usually discovered from the rating instrument with an item such as, "What grade do you expect to receive in this course?" (Note that Greenwald and Gillmore (1997a) found an average correlation of .43 across three samples. More on this below.) Cashin considers it a small relationship, but offers a few explanations for this phenomenon. It is possible that students who learn more get higher grades, and in turn rate the teaching effectiveness more highly than students who learn less. Or, it may be that instructors who grade easily, giving higher grades than are deserved, get "graded" highly in turn in the student ratings. There is some evidence that well-motivated students learn more, receive higher grades, and give higher ratings. Greenwald and Gillmore (1997a) argued that despite this correlation, student ratings do provide useful information, and it should simply be statistically corrected for. However, others have not agreed for a number of reasons (Kulik, 2001).

6) Several researchers have found that higher level courses receive higher student ratings than lower level courses. While these differences are small, where they appear, Cashin advises that courses at a given level should be compared with other courses at the same level.

7) Courses in some disciplines receive higher student ratings than others. The highest ratings occur in arts and humanities, then in descending order, biological and social sciences, business, computer science, math, engineering, and at the bottom, physical science (Ory, 2001, p.5). It is not clear why this is so, but among other possible explanations, Cashin proposes that teaching in these areas may differ, or that students may lack competence in quantitative reasoning for social science and especially for natural science. This may be another area requiring separate norms for different disciplines.

8) Instructors might expect that heavier workloads would result in lower student ratings. In fact, courses that have heavier work and intellectual demands show small but positive correlations with student ratings. Citing previous research with Sixbury, Cashin (1995) shows that student ratings correlated positively with amount of reading (.11), amount of other assignments (.16), difficulty of subject matter (.15), and how hard the student worked in the course (.29).

9) Student ratings that must be signed with the student's name tend to be higher than for unsigned ratings - presumably because anonymity permits more freedom of the student's real view.

10) Student ratings made with the instructor present in class tend to be higher than those with the instructor out of the room. Again, we may assume that absence of the instructor also permits students to feel freer to report truthfully.

11) If students are informed that their ratings will be used for decisions on promotion or tenure, the ratings tend to be higher. Cashin advises informing students of the ratings' purpose.

12) Majors tend to rate faculty higher than do non-majors.

13) Ratings administered during a final exam tend to get lower results than do ratings administered during a regular class.

14) Courses taken for general interest tend to get higher ratings than those taken to fulfill major or minor degree requirements.

There are two more findings, well-known to all researchers with human subjects, which affect the administration of student ratings as well:

15) There is a "halo effect," on student raters, namely the tendency for a general judgment of a person (positive or negative) to color or influence judgment of individual survey items. This may be a factor in the higher-ratings-if-instructor-present phenomenon.

16) Finally, the "error of central tendency," affects student ratings, namely peoples' inclination to avoid the extreme choices in a survey item, and to answer somewhere in the middle.

Limitations of students as raters. Felder and Brent (2004) notes that we must realize that students have limitations as raters. Specifically, students would not be likely to know whether a course is up to date, whether the assignments and tests are appropriately challenging, whether the course objectives are consistent with departmental goals (e.g. to serve as a prerequisite to another course, or serve some goal in the department's accreditation). Not all faculty will agree with these points, and the stated limitations may depend on the level, maturity, and experience of the students (recall that many U.S. students are older than the traditional age of 18-22 years).

Criticisms of the Use of Student Ratings

Readers might wonder, given the widespread acceptance of the reliability and validity of student ratings of instructors, why use of student evaluations is strongly criticized by some. Theall & Franklin (2001) indicate that "few issues in higher education are as sensitive, divisive, and political as faculty evaluation and in particular the quality and value of the information provided by students in their evaluations of teachers and courses" (p. 45). In some cases, faculty may be unaware of the literature, indeed even unaware of the basic statistical knowledge needed to interpret ratings correctly, as Theall & Franklin (2001) found in a 1989 survey (p. 46).

However, the criticisms seem to come from a number of voices, sometimes in non-mainstream venues, and they generally focus on validity and bias issues described above. The debate on the value of student ratings has moved from the pages of professional journals to university and organization newsletters, discipline-specific listservs, and self-publication on the web. For example, the Professional and Organizational Development Network (POD), the American Educational Research Association (AERA), and the comment pages of the Educational Policy Analysis Archives (EPAA) all have extensive threads on student ratings. These dissenting voices are sufficiently numerous, and the objections are sufficiently substantive, that they remind us that some serious controversy still surrounds the use of student ratings of instructors. On one side of the debate are those on the side of the validity studies; on the other side are those who actively seek contradictory evidence. Below is a survey of some of the major objections to the use of student evaluations of faculty from a variety of sources.

Non-validated ratings forms. One objection that has been made is that many university student ratings forms have not been validated. Commercially available instruments such as those mentioned above (SAIS, SIRS II, SEEQ, etc.) have undoubtedly undergone extensive validation processes, including the development of normative data across large numbers of subjects. However, none of the writings examined for this article assess how widely spread their use is. Many institutions, and units within institutions, develop their own student ratings forms, borrowing from still other forms. Ory and Ryan (2001) note that "it is fair to say that many of the forms used today have been developed from other existing forms without much thought to theory or construct domains. We have some evidence that traces the origin of several campus forms (including our own) to an item pool originally developed at the University of Michigan" (p. 32). Abrami, d'Appolonia, and Cohen found in their survey of 43 validity studies that 23 of the studies reviewed used student ratings instruments that were of local or unknown origin (p. 226). Why might institutions develop their own forms? Certainly a school may wish to tailor the items of their student rating forms so that they feel it accurately reflect their perspective on teaching effectiveness. However, such a form may or may not conform to the standards of validity. Trout (2000) points out that "validating a form is very expensive and therefore rarely done." He notes that his institution does

not use a validated form. Validation is also time-consuming. Certainly, use of non-validated forms calls into question the research arguing that student ratings are "moderately valid."

The grade and workload biases and the "dumbing down" of courses.

Probably the most controversial issue surrounding the use of student ratings of faculty is the knowledge that students who expect higher grades give higher ratings of their instructors. Centra (1993) noted that evidence for a relationship between grades and evaluations is inconsistent, but where observed, small. Commonly reported correlations are around .20.[2] In a recent paper, however, Greenwald and Gillmore (1997a), using a newly-developed rating instrument, examined data from 200 or more courses in each of three studies at the University of Washington. They found correlations between grades and student ratings of .48, .38, and .50 (a mean correlation of .45). The authors considered the theories that 1) Teaching effectiveness influences both grades and ratings; 2) Students' general academic motivation influences both grades and ratings; 3) Students' course-specific motivation influences both grades and ratings; 4) Students infer course quality and own ability from received grades; and 5) Students give high ratings in appreciation for lenient grading. In examining the data patterns closely, the authors conclude that only the final theory (the "leniency theory") provides the best explanation. Furthermore, in a separate publication, Greenwald and Gillmore (1997b) indicated that student ratings had consistent negative correlations with workload (that is, the more hours a courses required, the lower the student ratings, and vice versa). Recall that not all research on the grades-ratings relationship yields such strong correlations as the Greenwald and Gillmore studies. The authors state that it is not clear why their work at the University of Washington seems to contradict the findings of other research, and they hesitate to generalize the results beyond their own institution (1997b, p. 750).

However, as the reader may imagine, this and related findings have sent an alarm through the academic community. Trout (2000) argued that student ratings can encourage instructors to "dumb down their teaching." "The dynamic," he claimed, "is simple and widely understood, if seldom acknowledged. College instructors, as incentive driven as everybody else outside a Trappist monastery, are economic beings who calculate their self interest when making decisions affecting their income. If it takes consistently high evaluations scores to get raises, tenure, promotion and other perks, many instructors – consciously or unconsciously – will do what it takes to get those scores. That means that they will give students what they want – and many want lighter workloads, easier tests, and higher grades" (p. 3).

Stake concurs and describes his own experience of downgrading the academic level of his teaching: "I also believe that student evaluations can strongly influence the behavior of teachers, and for the worse. I changed my teaching dramatically because I was told by my Dean at the time that I had to keep the customers satisfied if I wanted to get tenure. (And I have not changed back since getting tenure.) I would not contend that the changes I made improved my teaching" (Stake, 1997). Lewis (1998) states, "If negative evaluations reflect student dissatisfaction with rigorous instructors and courses – and administrators use these evaluations in instructor retention decisions – then a natural selection of sorts may be watering down the quality of courses, a phenomenon dubbed 'higher education "lite."'" Not only can the need to keep students satisfied dilute the quality of education, but it could also discourage innovation in teaching methods (Lewis, 1998); instructors would simply be unwilling to risk their ratings.

In response to Greenwald and Gillmore's work (1997a, 1997b, and previous studies), and to various critics of student ratings, Marsh and Roche (2000) undertook studies (including two for the 2000 publication) that refute the grade leniency hypothesis. Reviewing their own and others' work, they note that the relation between grades and student ratings is consistently around .20, a comparatively small relation. They note also that "if a background variable has a similar influence on multiple measures of teaching effectiveness [student ratings, teacher self-evaluation, etc.], then the effect may reflect a valid influence on teaching effectiveness rather than a bias" (2000, p. 203). Re-analyzing Greenwald and

[2] Of course, a correlation is not evidence of a causal relationship, only of simultaneous co-appearance. However, many have responded to the implication that expectations of high grades somehow "cause" high student ratings.

Gillmore's data, and including students' perceived learning in their analysis, Marsh and Roche concluded that there is indeed a positive relationship between workload and student ratings, and that "classes expecting higher grades also report better perceived learning outcomes, and this accounts for the higher student ratings" (2000, p. 212). They then pursued a second, very large study with data from 5433 classes over a period of 12 years. Once again, they confirmed a correlation between grades and overall teaching evaluation of .198. (Another intriguing finding shows that grades correlate with some dimensions of the SEEQ, and not with others.) Further, they observed that when grades are above average, the relation between grades and student ratings is inconsistent (i.e. flat); only when grades are well below average is the relationship consistent and positive. It suggests an attributional explanation, namely that when people achieve successfully, they attribute their success to themselves; when people are not successful, they attribute their failure to causes outside themselves (e.g. the course or an instructor). Further, Marsh and Roche (2000) were able to confirm moderate relationships between workload and overall teacher ratings (.19) and overall course ratings (.25). The authors conclude that faculty cannot ensure high student ratings by offering lenient grading and low workloads. Instead, instructors should provide demanding and challenging materials and should encourage students to value their learning (2000, p. 226).

Legal considerations. A good deal of the disapproval of student ratings use expressed by faculty members has revolved around the impact of ratings on the quality of education, and the concern that their use leads to an infringement on academic freedom. The use of student ratings for administrative evaluation of faculty for personnel decision figures prominently in this issue.

In a four-part series of articles, Haskell (1997a) has examined this topic. Surveying a large number of faculty views, he found claims that student evaluations of faculty 1) are evidence of administrative intrusion into the classroom, 2) are used to intimidate faculty into politically conventional views, 3) encourage lowered teaching standards, 4) are responsible for considerable grade inflation, 5) prescribe classroom behavior 6) manipulate faculty behavior for promotion, salary, and other career-related decisions, 7) do not eliminate, but rather, encourage poor teaching, 8) promote "consumerism," 9) lead to the inappropriate dismissal of faculty, and ultimately, 10) comprise an unrecognized threat to academic freedom (1997a, p. 5).

Haskell exemplifies some of these concerns with case histories of faculty dismissed, not promoted, or reassigned teaching duties based (apparently) solely on student evaluations (p. 6ff). He points out that legal rulings are beginning to challenge the use of student ratings for personnel decisions. The issues include students being unaware of the purposes to which their ratings are put; the anonymous nature of the ratings (i.e. that raters are not held responsible if they misrepresent their course experience); the fact that ratings are essentially anecdotal and hearsay in nature; and the claims of ratings being invalid (1997a, p. 8). Haskell further points out that while students' grades must not be posted publicly with identifying information (such as social security numbers), student evaluations of faculty with anecdotal remarks are in some cases openly published by universities. He notes that the legality of this practice is beginning to be challenged (p. 9). He concludes: "[W]hat this article suggests is that SEF [student evaluations of faculty] is far from the benign instrument it may once have been in a more homogeneous political, gender, racial, and academically prepared student environment. Unfortunately, on many campuses the traditional model of student and teacher belongs to a past age. Faculty now teach in a litigious context. The new role and impact of SEF need to be reassessed accordingly" (1997a, p. 17).

Jeffrey Stake, law professor at Indiana University, disagrees with Haskell's argument that student ratings infringe on academic freedom. However, he expressly asserts his disapproval of student ratings. He states: "Almost anything that can be done to undermine the administrative practice of getting students to evaluate teaching ought to be done. One of my major concerns is that the process of asking students their opinions undermines the trust and faith they need to place in the teacher. Instead of saying, "Here is a great scholar and teacher; learn from her what you can," the administration of evaluation forms says to students, "We hired these teachers, but we are not sure they can teach or have taught you enough. Please tell us whether we guessed

right . . . " (Stake, 1997).

In still another critical review, Huemer cites a survey published in 1980 by Ryan and Birchler, in which 38% of professors survey admitted to "dumbing down" courses in response to student ratings. He cites also Sacks' *Generation X Goes to College* (1986), in which the author, threatened with losing tenure because of low student ratings, "became a teddy bear," offering anything and everything the students wanted (cited in Huemer, n. d).

Faculty perceptions of student ratings. Apart from legal issues, some faculty members object to student ratings for other reasons. In a survey of his institution's faculty published in a newsletter of the faculty senate, Birnbaum (1999) uncovered strong sentiments on student evaluations of teaching. Soliciting all faculty at the California State University at Fullerton, Birnbaum received 208 surveys in return. Below are several responses to survey questions on student ratings (1999):

If you were to RAISE standards for grades in your class, would it affect your student evaluations? Yes - 65.4%: No - 3.4% [the remainder were not reported].

If you were to INCREASE the amount of CONTENT (material) in your classes, would it affect student evaluations? Decrease student evaluations－65.9%; Increase student evaluations－4.8%.

Are student evaluations influenced by such variables as the teacher's personality, attractiveness, gender, race, dress, religion, ethnicity, sexual orientation, or disability status? Student ratings are unbiased－16.8%; Biased in favor of certain groups－52.4%; Biased against certain groups－26%.

Does the use of student evaluations encourage faculty to －"WATER DOWN"content in their courses? Yes－72.1%; No－26.9%.

Other findings included these: Nearly half of the responding faculty (48.6%) said they had decreased the amount of materials presented in class in recent years (14.9% said they had increased material, and the rest reported no change). While the majority reported no change in their grading standards, 32.3% indicated they had lowered grading standards, and 7.2% reported raising them (Birnbaum, 1999).

While we must recall that the Birnbaum faculty survey is a single study of a single institution, an internet search does show numerous other faculty voices expressing disapproval of the use of student rating in faculty newsletters, campus publications, internet discussions. The arguments contained in Haskell (1997) and Birnbaum (1999) appear to be representative of many.

Student ratings as measures of consumer satisfaction (rather than teaching effectiveness). Still another objection, while conceding that student ratings may indeed be valid, argues that 1) multi-section common exams of the type frequently cited in multisection validity studies generally measure low-level rote learning found in introductory classes (not higher-level outcomes such as critical thinking and problem solving); and 2) student ratings essentially measure students' satisfaction－that is, their affective reactions－rather than demonstrable cognitive learning.

Hake (2000), known for work in educational methods in physics teaching, argues that student ratings of teaching effectiveness defined as "customer satisfaction" can be useful, but that this must be separated from measures of actual cognitive gain. He cites the case where researchers "Halloun and Hestenes... found that four professors with good or superlative student evaluations, but who utilized traditional passive-student lectures, were almost totally ineffective in imparting any conceptual understanding of Newtonian mechanics to students in introductory physics courses at Arizona State University" (Hake, 2000). Hake also reminds us that McKeachie states that many multisection validity studies have used tests that assess factual knowledge, rather than, for example, reasoning, applications, problem-solving, etc. In such cases, "students' ratings of teaching effectiveness are likely to reflect a relatively unsophisticated conception of effectiveness" (McKeachie, 1997, p. 1219). Essentially, this means that in most cases, faculty who want to know if students have found higher-level teaching effective must design their own, non-validated forms, or rely on their exams and assignments. (Indeed, I advocate this below.) Hake (2002a) advocates high-quality standardized testing of both the cognitive and affective impacts of courses, and has been involved in developing such material for introductory physics courses.

The macro- vs. the micro-focus in student ratings forms. A final kind of objection to student ratings actually

hinges on the large amount of evidence for validity, e.g. in the multisection validity studies, in which the unit of analysis is the class, and mean ratings for classes. Of course, if we take the macro-analytic view which statistical evidence favors, looking across large collections of data, or even collections of whole studies, we lose the micro-analytic details. Marsh and Roche (2000), in their very large study, conclude by saying that future research on the grade-ratings relation must consider many more variables, for example, prior characteristics of students (e.g. year of study) and courses (e.g. required or elective), use of multidimensional ratings forms, teacher self-evaluations, grade leniency variables (variables which reflect grade leniency and variable which are direct measures), student achievement, and mastery. All this (not to mention the entire student ratings literature) implies profoundly complex field forces in the social setting of a classroom. Recall Fich's (2003) concern about variables for small populations (e.g. female computer science instructors) that may be lost in aggregate data.

Yet the research bias in the arena of student ratings of courses and faculty is toward to macro-analytic view. Bain [n.d.], in his overview of student ratings of instructors, identifies the generic question types that produce reliable and valid results (Provide an overall rating of the instructor; Give an overall rating of the course; Estimate how much you learned in the course; etc.), all rated on the ubiquitous 6-point scale, lowest to highest. He notes that questions about teaching processes (Did the instructor communicate well? Was the instructor available and willing to provide assistance outside the classroom? etc.) may help instructors improve teaching, but notes that they may be misleading because they assume a homogeneity that may not reflect reality. That is, if an instructor achieves results with an unorthodox method, or teaches in a field where the conventional methods are not used, the ratings on these process items may be low; however, the results may be good, as would be indicated with the results-oriented questions. Thus the generic results-type questions satisfy the demands of statistical research, reflect student satisfaction, and thereby please administrators - however, they are minimally informative to the faculty member looking for substantive understanding of what is transpiring in his or her classroom.

Faculty determined to investigate what happens below the bland appearance of satisfaction might be surprised to find out what is going on their student ratings' process. A recent study used verbal protocol analysis (think aloud interviews) to examine what a group of medical students were thinking as they completed a conventional course evaluation form for a science course (Billings-Gagliardi, Barrett, and Mazor, 2004). Students were asked how they interpreted the items, how they formed judgments, and how they selected the responses as they completed a typical student rating form. Students interpreted educational terms differently from one another and differently from common usage. Sometimes students made judgments on unique rather than expected criteria. Judgments were sometimes tempered by students' perception of the faculty member's caring or effort. Additionally, students avoided the lower end of the scale, used the highest choice selectively, and the second highest choice indiscriminately. The authors concluded that their results challenge fundamental assumptions about the student rating process, particularly assumptions about how well students understand what they are reading on the forms, and whether students' ratings indeed measure teaching effectiveness.

These are, of course, a few single voices which contrast with the systematic literature of, for example, the large question of validity. However, a look at both the formal and informal arguments for and against conventional student ratings forms suggest that there are weaknesses inherent in blind and wholesale adherence to student ratings as measures of teaching effectiveness. At least, there are a sufficient number of issues to warrant caution in the acceptance of student ratings as the sole indicator of teaching effectiveness. Additionally, these issues remind us that student ratings of instructors are administered in a profoundly complex setting of social field forces, and that they reflect these field forces incompletely.

Best Practices for Institutions

In this section I have collected the "best practices" for student evaluation of faculty and courses in the U.S. setting. Best practices, as they can be gleaned from the literature, frequently address both student evaluation of faculty as used for both *formative purposes* (improving teaching effectiveness) and *summative purposes*

(evaluating faculty for personnel decisions). While we do not see the implementation of evaluation for both purposes in Japan, many of the recommendations addressing summative evaluations have useful implications for formative evaluation as well. This section contains a sample of authors' best advice:

- One best practice recommended by virtually all authors addressing teaching evaluation (Cashin, 1996; Centra, 1993; Fich, 2003; Theall and Franklin, 2001; McKeachie, 1997) is that student ratings should never be used by themselves, but rather complemented by other evaluation measures (peer observation, developmentof a teaching portfolio, experts' observation, etc.).
- d'Apollonia and Abrami (1997) and Cashin (1995) point out that many authors agree that when student ratings are used for formative purposes, specific items are most helpful. A number of researchers feel that when student ratings are used for summative purposes, the average of a few global measurements should be sufficient. However, some feel that because teaching is a multifaceted activity, specific scores reflecting the multidimensionality of teaching are more suitable (d'Apollonia and Abrami, 1997, p. 1203).

Cashin, in his article "Developing an Effective Faculty Evaluation System" (1996) enumerates a number of good general evaluation practices. Many of these are also echoed in Theall & Franklin (2001). The following is a distillation of their best practices.

- Establish the purpose of the student ratings before they are implemented. Whether used for formative or summative purposes, identify what data will be gathered and how it will be analyzed, assembled and used. Ensure not only that institutional goals are addressed, but also the goals of individual units and departments—which may be different from one another. The reason for making the uses of data collection clear before it is done is, Cashin says, justice. "It is unfair to collect data without everyone knowing who will received what information for what purposes, because such knowledge can influence the responses of students and of others" (1996, p. 2). In particular, faculty who understand the purposes of evaluation will be more likely to accept it and less likely to mount legal challenges.
- Where evaluation data is used for improvement of faculty teaching effectiveness, ensure that systematic help is available in the form of instructional consultants or master teachers. Theall and Franklin (2001) note that evaluation without such support is punitive. Often, on U.S. campuses, department heads consult with individual faculty on their teaching. Cashin points out that this can be very helpful, but that department heads are involved in personnel decisions, and that could be a conflict of interest.
- Participants in the evaluation process should be *significantly* involved in the decision-making. The goal is to maximize a sense of acceptance and ownership. Extensive, open communication should be maintained before, during, and after the implementation of an evaluation system. [Development or] change in an evaluation system involves more than "having a good idea," Cashin points out. It is about changing attitudes, values, traditions, and their attendant emotions, i.e. an affective process. Approaching faculty with a strictly empirical-rationalist strategy, or a power-coercive strategy to force change on them, Cashin says, is counter-productive.
- Ensure that the evaluation system is flexible, i.e. that particular departments, divisions, or units are fairly addressed in evaluation forms. Acceptable teaching, research, and service practices differ in different disciplines—this information must form a context for the interpretation of evaluations. These differences must be spelled out to forestall unfair comparison across fields. (Recall that research shows that some disciplines get higher ratings than others.)
- Ensure that the evaluation system is legal. (Since this is a complex area, Cashin refers the reader to appropriate sources for the U.S. context.)
- Make faculty responsibilities which are being evaluated clear at the outset, and determine their weighting or priority. Tradition faculty responsibilities in the U.S. are teaching, research, and departmental/institutional service. Some fields may, Cashin notes, involve a unique area of responsibility, e.g. student supervision in the areas of nursing or education. Two new areas for evaluation are beginning to appear in the literature: Professional competence (degrees, licenses, experience)

and professional behavior (e.g. ethical behavior related to research and teaching, collegiality). Cashin comments that it is relatively difficult to define professional competence, and expresses the concern that in practice, only negative behaviors are likely to be identified in faculty evaluation.

- Ensure that data-gathering methods are technically reliable and valid.
- Train evaluators to evaluate (that is, train students for student ratings, train faculty and administrators for peer observation or portfolio evaluation, etc.). Train supervisors to give feedback in a constructive way.
- Maintain appropriate confidentiality. "Faculty and administrators should take such confidentiality as a very serious professional and ethical obligation," Cashin advises. [In the U.S., the instructors are routinely required to leave the room before the students begin their ratings. They do not see or handle the evaluation results until grades have been submitted.]
- Reward effective performance. An evaluation system will not be effective if it has no impact on faculty behavior: Effective teaching (and other performance) must be accurately recognized and rewarded (and must be seen by faculty as being accurately recognized and rewarded).

Cashin concludes that developing a faculty evaluation system (including student ratings) is a time-consuming process. Interested readers who wish to see a formal, technically detailed list of advice generated from the experience of Kansas State University's IDEA Center are referred to Cashin (1990).

Felder (1993) offers more specific suggestions on the administration of student ratings for formative purposes. His brief, practical set of ideas is quoted in full, with editing.

- ***Collect overall course-end ratings of instruction.*** For example, "Rate the instruction you received in this course on a scale from 1 to 5, with 5 being the highest response." Ratings of this sort are most effective when the numbers on the response scale are clearly defined. [Rather than] "excellent," "above average," "fair," etc., ...you can get greater discrimination with a variation of the following instruction: *When responding, use as a basis of comparison all of your previous high school and college teachers. A response of 5 denotes one of the three or four best you've ever had; 4 = top 25%; 3 = 40-75%; 2 = bottom 40%; and 1 = one of the three or four worst you have ever had.* An instructor whose average rating is close to 5 on this scale is clearly doing a superb job and deserves nomination for an outstanding teacher award, and serious problems obviously exist if an instructor's rating is consistently close to 1. Ratings close to 4 indicate commendable teaching performance and ratings close to 2 suggest the need for corrective measures.
- ***Collect ratings of individual aspects of instruction.*** To get the most out of a course-end evaluation, supplement the overall rating with ratings of specific aspects of teaching performance, such as clearly stating expectations, providing frequent examples, repeating difficult ideas, pointing out practical applications, answering questions thoroughly, preparing tests that reflect course content and emphasis, etc. (General questions about the instructor's preparedness and knowledge of the subject tend to be less useful.) The responses help identify areas of weakness and may provide ideas about how to improve teaching in the next course.
- ***Administer and collect ratings in the same session.*** To be sure that the evaluations reflect a true cross-section of student opinion, administer and collect them in a single class session rather than counting on students to return them later. Results of evaluations for which the return rate is less than a minimal percentage should be regarded with deep suspicion: The recommended minimum is 50% (classes of 100 or more), 66% (50-100), 75% (20-50), and 80% (<20).
- ***Collect evaluations midway through a course rather than waiting until the end.*** If the goal is to correct teaching problems and not just to identify them, find out what the problems are while enough time remains to do something about them. Ask open-ended questions on midcourse evaluations, leaving plenty of space for the responses: *1. What do you like best about this course and/or the instructor? (List up to three things.) 2. What do you like least about the*

course and/or instructor? (List up to three things.) 3. If you were the instructor, what would you do to improve the course? Felder emphasizes that mid-term feedback and instructional consultations can pay off. He summarizes a well-known study on instructional effectiveness by Cohen (1980): "In one study, instructors who received no feedback in the first half of a course received average end-of-term ratings in the 50th percentile of the population studied; instructors who received feedback scored in the 58th percentile; and instructors who got both feedback and instructional consultation scored in the 74th percentile" (Felder, 1993, p. 3)

- ***Collect evaluations from small groups rather than from every student.*** One problem with individual evaluations is that many of the responses may reflect isolated gripes rather than widely-held opinions. Another is that students may be fearful of offering negative criticism while a course is still in progress, even if the evaluations are anonymous (as they should be). A good way to counter both of these problems is to collect evaluations from groups of four or five students rather than from individuals. The students in a group should spend 5-10 minutes discussing the three questions given above, then prepare a collective evaluation that only includes points agreed upon by several group members.
- ***Interview student representatives.*** Designate certain students as representatives of subgroups within the class. At one or more times during the semester, meet (or ask a colleague to meet) with the representatives to share the concerns of their constituents and to discuss possible measures to correct perceived problems. This procedure tends to generate constructive criticism at a level rarely attained through written evaluations and also gives students a greater sense that their opinions are valued.
- ***Use a variety of sources of feedback.*** Collect retrospective teaching evaluations from graduating seniors and alumni. Have faculty colleagues observe your teaching and provide feedback. Have one of your classes videotaped and review the tape (brace yourself–you may not be thrilled by everything you see).
- ***Work with an instructional consultant to interpret student feedback and plan teaching improvement strategies.*** It is one thing to know that some students consider you a poor lecturer or think your tests are unfair and quite another to know what to do about it. Many universities have instructional consultants whose job is to help faculty members improve their teaching. These people can provide a variety of services, such as helping design and administer evaluation questionnaires, interviewing classes or groups of students about their perceptions of the instruction, observing and critiquing live or videotaped lectures, and working with instructors to help interpret evaluations and plan corrective strategies. If no one like this is available on your campus, ask a faculty colleague with a reputation as an outstanding teacher to work with you. (Felder, 1993).

Fich (2003) offers still other useful advice on the use of student ratings information (again, the suggestions are quoted, with editing):

- Only compare results from similar courses, including such factors as class size, course level, difficulty of the material, whether the course is theoretical or applied, and whether the course is required. (Recall the research on biasing variables above.)
- Avoid subjective items, like "overall effectiveness of instructor," which are vulnerable to bias, even though administrators want a single summary number.
- [Be aware that] results of student evaluation have low precision. Multi-section courses, even though graded by the same instructor, may show significant differences in supposedly objective items such as "assignments are graded fairly" or "Returns work promptly." Given the large number of possible sources of error that can arise when students evaluate teaching, it is only reasonable to interpret the results using a very coarse scale: outstanding, good, or poor. (Fich notes that at least one science faculty give students a comprehensive list of possible positive and negative course characteristics, and asks them to choose from the list, rather than rate aspects on a scale.)
- Eliminate inappropriate student comments, including slanderous, abusive, or personal remarks. University staff administering the evaluations should remove

forms containing biased or offensive comments and not include their ratings in any compilations. A better way to discourage inappropriate or unfair ratings is to enable staff (but not faculty) to identify the rater's identity. (Fich, 2003)

Ory (2001) offers this useful suggestion:

- Develop separate norms for required and elective courses.

Bain (n.d.) advises the following kinds of questions for a student ratings form:

- Student ratings forms should contain questions asking for the following: An overall rating of an instructor; an overall rating of a course; an estimate of how much the rater learned in the course; the effectiveness of the instructor in stimulating interest in the subject; the degree of intellectual challenge of the course. Additionally, the following kinds of information should be gathered: The student's year in college; the student's major; whether the course was required for a major credit, a minor credit, or was an elective, or was taken for general interest; the level of the student's interest in the subject before taking the course. Questions about the teaching process will be helpful as well: Did the instructor communicate well? Was the instructor available to help students outside the class? Was the course well-organized? Instructors should add their own questions about particular aspects of the course.

Institutions use these best practices to greater or lesser degrees. In cases where the practices are soundly and wisely done, the consequences are good. In cases where student ratings are carelessly administered, or politicized inappropriately, the consequences are less than desirable. It is worth noting what these consequences are. Ory, citing work with Ryan (2001) identifies intended and unintended consequences of student ratings. It can immediately be seen that these consequences comprise a kind of "Do's and Don'ts" list:

Intended Consequences of Student Ratings Use

- Instructors are rewarded for having excellent rating results (e.g. salary, promotion, awards, recognition).
- Instructors with very low ratings are encouraged by their departments to seek help, possibly from colleagues or a campus faculty development office.
- Students perceive and use ratings as a means for indicating suggestions for improvement.
- Students have more information on which to base their course selection.
- Instructors use ratings as motivation to improve their teaching.
- Students perceive ratings as a vehicle for change.

Unintended Consequence of Student Ratings Use

- Instructors alter their teaching in order to receive high ratings (lower content difficulty, provide less content, give only high grades).
- The campus rewards poor teaching (lowering faculty standards).
- Due to their convenience, the campus looks to student rating as the only measure of teaching quality.
- The content of the student rating form may determine what is addressed in the classroom.
- Students reward poor teaching by believing they can give high ratings in return for high grades.
- Ratings are used to make discriminations between instructors that cannot be supported by the data.
- Due to the high stakes involved, instructors fail to follow proper administration procedures.
- The ratings process becomes a meaningless activity that is only performed by students and instructors because it is mandated. (Ory, 2001, p. 10-11)

Complementary Alternatives to Conventional Student Ratings Forms - Finding Out What We Really Want to Know

Student ratings of courses and faculty performance are undeniably useful. However, they are not perfect. There are strong reasons to retain the careful and discriminating use of conventional student ratings forms. At the same time, the foregoing review does suggest that - at the least - faculty instructors and administrators should also use alternative and complementary evaluation strategies. A number of these alternative strategies are suggested in the Best Practices section above - peer observations, small groups rather than individual responses, student focus groups (lead by a trained leader), and the like.

In this section, I want to go beyond Bain's suggestion that faculty should add a few questions to their ratings form. I propose that in addition to using the generic ratings items designed to fit all cases, faculty members should write their own complementary alternative evaluation forms. These should incorporate thoughtful items that ask students to give direct and specific feedback about classroom events and activities, and their learning experience (i.e. the way in which they find themselves engaged in the coursework). While lacking the conventional claim of validity, such forms would provide substantive and real information about how students are experiencing the course. I arrived that this conclusion some years ago after struggling with what seemed to be inappropriate components of conventional forms, and confusing results from the students' responses. To elaborate, I must ask the reader's indulgence as I detail some of my history with conventional student ratings forms, and how I came to these alternatives.

My collection of student evaluation forms goes back to 1983.[3] At that time, my main teaching load was in English as a second language (for international students), and our evaluation form was a slightly simplified version of the form used for native speaking students at my college. This form bore a distinct resemblance to Marsh's 1976 SEEQ, including items on teaching, course organization, instructor/group interactions, instructor/individual interaction, and general evaluation. (However, there were no item clusters on learning, instructor enthusiasm, breadth of content, assignments, exams, or student and course characteristics, as there are in the SEEQ.)

While many of the ratings form items made reasonable sense (e.g. My instructor explained course purposes, My instructor answered the student's questions, etc.), one can only wonder if intermediate-level non-native English students were able to judge whether the instructor "understood the subject he/she was teaching." In language skills classes, employing a specific range of standard language teaching techniques for listening, reading, writing, etc., how could students interpret the statement "My instructor used several different teaching methods to present the course material" ? In an intermediate Listening class, what might students make of "My instructor helped students to increase their interest in the subject" ? These were items suitable to content teaching rather than skills teaching.

The difficulties interpreting student ratings did not diminish when I moved to teaching (mostly native speaker) freshman composition and senior thesis courses, where students produced a group research paper. The items about "exams with instructional value" and "fair tests of course material" were irrelevant, since course grades were given for written papers, and exams were not given. Nor did the "good textbooks" item suit the senior thesis course, which had no textbook. The item about whether an instructor "presented origins of concepts" - suitable to a history, economics, or biology course - were simply irrelevant. Sometimes students appropriately chose the "Not applicable" option, but many chose 3 - "uncertain" on the ubiquitous 5-point scale, or some other number. Whatever number they chose affected the average rating, of course, sometimes pulling it downward, and added to the lack of interpretability. One brave soul in the senior thesis class bluntly wrote in the comment space, "This evaluation doesn't apply."

Even for items which did make sense for the kind of classes I taught, student responses were sometimes difficult to understand. Why did one student say the class had improved his/her English and learning skills "very much" ; and then rate the class and the teacher's overall performance as "3 - average" ? Or what did it mean when a student rated the instructor "excellent" and rated the course "average" ?

I cannot claim that these evaluations were entirely unhelpful. When I started a freshman composition program with 8 assigned papers to be completed, student evaluations let me know in no uncertain terms that it was too much work, and I retrenched to a more reasonable 5 assigned papers. Still, too much of the ratings data was uninterpretable. As Centra pointed out, "One reason the so-called diagnostic items on most student ratings may not lead to greater changes in teaching is that they are not specific enough" (1993, p.84).

[3] Ratings forms that I had collected from 1973-83 were lost in a shipping mishap.

Finally, in 1986 I began to wise up. I began to design student ratings forms that asked specifically what I wanted to know about particular events in the students' experience of the course. I isolated five or six key outcomes - skills, abilities, areas of knowledge - that I wanted to students to leave a course with, and I wrote statements for each outcome that allowed the students to indicate how their skills and abilities had changed from the beginning to the end of the course. Instead of the usual vague "very poor - poor - average - good - excellent" descriptors, I unwittingly followed Felder's (1993) and Bain's (n.d.) advice and made specific descriptors to quantify the skill or ability level:

1 - none at all
2 - some, but not enough for courses
3 - enough for this course
4 - enough for other college classes
5 - professional ability for jobs

Then I asked students to tell me about their progress with specific skills, abilities, and understanding. Thus, students in a writing class could tell me, using the appropriate number:

My ability to organize my thoughts in writing went from (e.g. 2) to (e.g. 3).
My awareness of different types of writing in academic work, and the different styles appropriate for them, went from ___ to ___.

Or, in a listening class, students could tell me:

My ability to grasp the main idea of a lecture went from ___ to ___.
My ability to take effective lecture notes went from ___ to ___.

Or in a reading class, students could indicate:

My understanding of paragraph organization went from ___ to ___.
My ability to predict what comes next in reading went from ___ to ___.

(More recently, this particular strategy is being used in ongoing research on the perception of students' English skills development; see Winskowski and Hanna, 2003, for a preliminary report.)

Additionally, I wanted to know specifically how students found particular components of a course. For example, after developing an extensive self-guided library workbook for the students in a composition course, I incorporated this into the students' course evaluation:

Check whatever is true for you.
When doing the library exercises I found that:

___ I enjoyed exploring all the books [indicated in the workbook exercises]
___ I didn't like having to look at all the books

___ I knew most of the information in the exercises
___ I knew some of the information in the exercises, but learned a lot about library systems
___ a good deal of the information in the library workbook was new to me

___ the quizzes were reasonable tests of the information in the workbook
___ the quizzes were not very good tests of the information in the workbook

___ now I feel confident about finding information from reference books, indexes, and periodicals
___ now I'm still uncertain whether I can successfully find information from reference books, indexes, and periodicals

For each aspect of the class activity (the paper-revision process, computerized quizzes, etc.), I tried to think of all possible alternative answers (within reason) so that the student could simply find the case that fit and tick it off.

I administered my "home-made" student rating form along side the official college rating form. In addition to quite a lot of rather vague information about whether I as the instructor "explained course material clearly, was open to students' viewpoints, was well prepared and organized for class," etc., I received a wealth of information about specific writing skills the student were working on, how they found the computer-administration of quizzes, and what they were learning about the paper-writing process (journal-writing, peer feedback, revision). Additionally, students seemed to be encouraged to write extensive, paragraph-long comments on what worked well for them and what did not, and why.

Encouraged by what I was learning about the courses

from students' perceptions, I continued to experiment with alternatives to conventional student ratings forms. I tried short-answer items, such as these, taken from both mid-term and end-of-term from an advanced composition class and a linguistics class:

- Rank the readings, from best to worst, and add any comments (useful? engaging? informative? or ... ?)
- What was the result of your reading logs? Was there any effect on the readings for your other courses? Would you change the design of your reading logs? If so, how?
- Are you learning a lot of linguistics?
- What is the most interesting topic [in the course]? The least interesting topic?
- What do you feel in this course will be most valuable to you for your major?
- How do you like the research project you are working on?
- What has been the most difficult part of the course?

Recently, in trying to determine what activities students in a language class found helpful, I tried a different approach to feedback at midterm. I listed all the different activities we were doing, and asked students to rate their helpfulness:

Please put a number by the following class activities to show how helpful for learning English each one is. 1 = very helpful, 2 = somewhat helpful, 3 = not so helpful

_____ watching the video episodes on my own
_____ watching video episodes in class with some explanation from the teacher
_____ exercises from the Video Book
_____ exercises from the Conversation Book for two people
_____ exercises from the Conversation Book for three-four people
_____ practicing dialogue from the video episodes (first reading, then saying)
_____ learning about collocations in the dialogue
_____ exercises on pronunciation, intonation, and word stress
_____ exercises from Conversation Strategies (the green book)
_____ audiotapes
_____ email conversations
_____ other (please explain)

In a few cases, the students found some activities more helpful than I would have guessed. Other activities that I thought were important were not seen that way by the students.

Most recently, in offering a content course (Comparison of Japan and the U.S.), which involved a good deal of lecturing in English to non-native English speaking students, I began to include, on each week's study questions, the following items:

- What percentage of the lecture did you understand today?
- What questions [about the lecture] do you have?

Responses to the first question showed that students' understanding ranged from an astonishing 20% to 90% (with a mean of 52%). It was a wake-up call, making me realize that I had to provide vocabulary ahead of the lecture, ensure that whiteboard notes were substantive and redundant to my lecture, provide time for students to work on study questions in class, so they could help each other, and wander the aisles while they worked to answer questions. This also allowed me to pitch my lecture delivery at least a little more realistically than I might have otherwise.

Finally, as Bain (n.d.) suggests, I have even taken the conventional format of student ratings forms, with conventional items ("The course was well-organized;" "I learned a lot from this course;" etc.), and added questions about the specific outcomes of the courses (in this case Seminar on Japan and the U.S., and Business English) that I was aiming for:

- I understand U.S. American culture better after being in this course.
- This class made me see the Japan differently from before.
- When I completed the study questions, it helped me to understand what was important.
- I learned a lot about English-style business letters.
- I learned a lot about English-style business memos.
- The textbook was useful for colloquial English expressions.
- I learned some practical knowledge about writing (e.g. about punctuation, capitalizing, etc.).

- I regularly used the listening tapes that went with the textbook.

It can be seen that complementary course evaluations can take many forms. Indeed, readers can see that with a little reflection on one's teaching goals, it is relatively easy to generate feedback. The key question driving the generation of items is: What knowledge, skills and abilities, and understanding do I want the students leaving the course with? Once this is clear, items can be made accordingly.

As I look back through 20 years' worth of student evaluations, I realize that I unwittingly incorporated Hake's admonition (2002a) that students' cognitive evaluation must be solicited in addition to with affective evaluation. I felt that students' evaluation of a course and students' perception of their own performance are often two sides of the same coin. While my alternatives did not specifically ask the students how they saw my effectiveness as an instructor (it did not occur to me to couch my inquiries in these terms), they allowed me, in various ways, to ask students specifically what was happening to them in their learning experience, how their skills were changing, how they saw these changes, and what elements of the courses were causally related to these effects, from their perspective.

I confess I did not consistently develop or use these alternative evaluation forms, nor could I always systematically apply what students were telling me to refining courses, due to heavy teaching loads. However, through the years, I have developed a portfolio of complementary evaluation forms which I can quickly adapt to new courses.

It is my hope that other instructors will be encouraged to develop their own complementary evaluation forms, designed to answer specifically and directly what they need to know about how their courses are working. In a recent edited volume called *Techniques and Strategies for Interpreting Student Evaluations* (2001), W.L. Rando offers three guidelines for faculty who are writing their own questions. First, he confirms that questions should be based on specific objectives for student learning. "Learning objectives are precise statements of the changes we are trying to create in our students. When we ask ourselves how we expect students to be different as a result of our teaching, we can construct questions that truly speak to our core intentions and then assess the effects of the methods we choose" (p. 79-80). Further, Rando goes on to advise writing items that encourage students to think deeply and reflectively about their own learning. Asking specific questions about how an assignment changed students' thinking furthers their understanding, enlightens their awareness, and trains students to be witnesses to their cognitive processes. Finally, Rando encourages faculty to ask specific questions of students about which elements of the course teaching methods further the students learning and fit the course and departmental objectives, and which do not (Rando, 2001).

In the same volume, Svinicki points out that students don't often see good examples of helpful written course feedback, and advises us to give students written models. The models should show feedback that is specific, that concentrates on observable behavior (avoiding personal comments or emotionally charged comments), and provide alternatives to criticized behavior (2001, p. 20-21). She adds that it is important to give students multiple opportunities to give feedback, to give them time to think over and prepare their comments (p. 22-23). Echoing Cashin (see Best Practices), she is advising us to carefully train the students in their observational skills, as we would in any other area.

A final recommendation from this volume comes from Lewis (2001), who notes that many of us have received, from the same class, one student comment that seems radically positive, and another radically negative, leading us to possibly dismiss this feedback as unreliable. He advises that we can essentially do a kind of content analysis: Group the comments according to the overall ranking of the course (e.g. one of the best, above average, average, below average, and far below average). Add a second component to the groupings to classify the content of the comment - subject matter (or analytic/synthetic approach to the topic), organization/clarity, instructor-group interaction, instructor-individual interaction, and dynamism/enthusiasm. Locating the comments in the resulting matrix can bring order to them and help the

instructor locate where student concerns are.

Conclusions

In the U.S., the initial purpose for student ratings forms for measure teaching effectiveness has expanded to now include assessment for promotion, tenure, and related evaluation of faculty. It is an empirical question whether universities in Japan will adopt this practice. If they do, they would have the benefit of being able to observe how this practice has unfolded elsewhere.

From data on student ratings in the west, it is not clear－cut whether ratings instruments measure independent dimensions of teaching, or one global dimension. Could the dimensionality of teaching be determined for Japanese university classrooms? It is an empirical question and would be a useful focus of exploration.

It is considered that in the literature reviewed, across large groups of people, students' ratings of faculty are reliable and moderately valid. It is an empirical issue whether we would find a similar reliability and validity in ratings in Japan, and this would also prove a fruitful initial area of research.

University classrooms have tremendously complex field forces, and there are certainly strong indications that student ratings data is more complex than we know. We know that many things unrelated to teaching effectiveness can influence student ratings, among them expected grades, the course' discipline, the students' year in school, whether the course is required or an elective, the physical presence of the instructor, and so on. There may even be possible connections between faculty personality and type of course taught. The variability in students' responses indicated in think-aloud protocols surely confirms that there is more going on than we know. It may be presumed we would find some corresponding complexities in the Japanese higher educational setting. For example, some preliminary research in one Japanese university indicates that small, elective classes held for third- and fourth-year students get higher ratings than large, required ones held for first- and second-year students (Ruthven-Stuart, 2004). The particular factors (class size, required status, etc.) remain to be disentangled. There certainly will be unique elements in the Japanese setting, however. One colleague notes that in Japan the large number of university language instructors who are trained in literature, history, and other fields may get different (lower) evaluations than those trained in language teaching.[4] Ultimately, the pattern of these complexities in the Japanese university setting is an empirical question and is yet to be determined.

In the U.S., because of the impacts of student ratings on faculty status, there are beginning to be legal challenges to their use and the use of ratings results. We see no such issues in Japan at present; whether future legal issues might arise remains to be seen. Further, student ratings remain a topic that is controversial. Will university faculty in Japan come to accept - or debate - the use of student ratings?

One sure conclusion that can be drawn is that student ratings of teaching effectiveness (or whatever it is that they measure) seem to be a permanent and growing part of the educational landscape. Imperfect as they are, student ratings of faculty and courses are a "window" through which faculty can investigate their own teaching effectiveness. In Japan, the nature of education is changing. Opening admission to non-traditional-age population, to foreign students and to other groups could certainly necessitate revision of course and program design, and feedback from students is a natural next step. To the extent that faculty are willing to examine their own practices and processes, they can exert control over their effectiveness in the classroom.

[4] Ryoko Kurihara, personal communication, November 27, 2004.

References

Abrami, P.C., d' Apollonia, S. and Cohen, P.A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 219-231.

Adams, J.V. (1997). Student evaluations: The ratings game. *Inquiry, 1*(2), Fall, Virginia Community College System.

Bain, K.R. (n.d.) Using student ratings. New York University Center for Teaching Excellence. Online at: http://www.nyu.edu/cte/white.html#Number2. Accessed on Oct. 18, 2004.

Billings-Gagliardi, S., Barrett, S.V. and Mazor, K.M. (2004). Interpreting course evaluation results: Insight from think aloud interviews with medical students. *Medical Education, 38*(10), 1061 - 1070.

Birnbaum, M.H. (1999). A survey of faculty opinions concerning student evaluations of teaching. *California State University, Fullerton Senate Newsletter*, Vol. XIV, No. 1, Fall. Online at: http://faculty/fillerton.edu/senatenews/page2.html.

Cashin, W.E. (1990). Student ratings of teaching: Recommendations for use. IDEA Paper No. 22, Kansas State University Center for Faculty Evaluation and Development. Online at: http://www.idea.ksu.edu/products/Papers.html.

Cashin, W.E. (1995). Student ratings of teaching: The research revisited. IDEA Paper No. 32, Kansas State University Center for Faculty Evaluation and Development. Online at: http://www.idea.ksu.edu/products/Papers.html.

Cashin, W.E. (1996). Developing an effective faculty evaluation system. IDEA Paper No. 33, Kansas State University Center for Faculty Evaluation and Development. Online at: http://www.idea.ksu.edu/products/Papers.html.

Centra, J.A. (1993). *Reflective Faculty Evaluation*. San Francisco: Jossey-Bass.

d'Apollonia, S. and Abrami, P.C. (1997). Navigating student ratings of instruction. *American Psychologist, 52,* 1198-1208.

Felder, R.M. (1992). What do they know anyway? *Chemical Engineering Education, 26*(3), 134-5 (Summer). Online at: http://www.ncsu.edu/felder-public/Columns/Evals.html.

Felder, R.M. (1993). What do they know anyway? II. Making evaluations effective. *Chemical Engineering Education, 27*(1), 28-29 (Winter). Online at: http://www.ncsu.edu/felder-public/Columns/Eval2.html.

Felder, R.M. and Brent, R. (2004). How to evaluate teaching. *Chemical Engineering Education, 38*(3), 200-202.

Fich, F.E. (2003). Are student evaluations of teaching fair? *Computing Research News*, Vol. 15, No. 3. Online at: http://www.cra.org/CRN/articles/may03/fisch.html

Greenwald, A.G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52,* 1182-1186.

Greenwald, A.G. and Gillmore, G.M. (1997a). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52,* 1209-1217.

Greenwald, A.G. and Gillmore, G.M. (1997b). No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology, 89*(4), 743-751.

Hake R.R. (2000). Student evaluations (was "Re: On Correlations") PhysLrnR post of 6 July. Online at: http://listserv.boisestate.edu/cgi-bin/wa?A2=ind0007&L=physlrnr&P=R553&X=42456A068128359A91&Y=rrhake@earthlink.net.

Hake, R.R. (2002a). Lessons from the physics education reform effort. *Conservation Ecology, 5*(2): 28. Online at: http://www.consecol.org/Vol5/iss2/art28.

Hake, R.R. (2002b). Problems with student evaluations: Is

assessment the remedy? Post to POD of 25 Apr. Online at: http://www.stu.ca/~hunt/hake.htm.

Haskell, R.E. (1997a). Academic freedom, tenure, and student evaluations of faculty: Galloping polls in the 21st century. *Educational Policy Analysis Archives, 5*(6). Online at: http://olam.ed.asu.edu/epaa/v5n6.html.

Haskell, R.E. (1997b). Analysis and implications of views from the court in relation to academic freedom, standards, and quality instruction, *Educational Policy Analysis Archives, 5*(21). Online at: http://olam.ed.asu.edu/epaa/v5n21.html.

Huemer, M. (n.d.). Student evaluations: A critical review. Online at: http://home.sprynet.com/~owl11/sef.htm. Accessed 08/29/04.

IDEA Center, Kansas State University Homepage. Online at: http://www.idea.ksu.edu/. Accessed 12/01/04.

Kulik, J.A. (2001). Student ratings: Validity, utility, and controversy. In M. Theall, P.C. Abrami, and L.A. Mets, (Eds.). *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* New Directions for Teaching and Learning, No. 109, San Francisco: Jossey-Bass.

Lewis, K. G. (2001). Making sense of student written comments. In K.G. Lewis (Ed.). *Techniques and Strategies for Interpreting Student Evaluations.* New Directions for Teaching and Learning, No. 87. San Francisco: Jossey-Bass.

Lewis, R. (1998). Student evaluations: Widespread and controversial. *The Scientist,* 12(9):12, Apr. 27. Online at: http://www.the-scientist.com/yr1998/apr/prof_980427.html.

McKeachie W.J. (1997). Student ratings: The validity of use. *American Psychologist, 52.* 1218-1225.

Marsh, H.W. (1976). Student evaluation of educational quality (SEEQ). Center for Excellence in Learning and Teaching, Pennsylvania State University. Online at: http://www.psu.edu/celt/SEEQ.html.

Marsh, H.W. and Roche, L.A.. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist,* 52, 1187-1197.

Marsh, H.W. and Roche, L.A.. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology, 92*(1), 202-228.

Murray, H.G., Rushton, J.P. and Paunonen, S.V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82*(2), 250-261.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. In K.G. Lewis (Ed.). *Techniques and Strategies for Interpreting Student Evaluations.* New Directions for Teaching and Learning, No. 87. San Francisco: Jossey-Bass.

Ory, J.C. and Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P.C. Abrami, and L.A. Mets, (Eds.). *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* New Directions for Teaching and Learning, No. 109, San Francisco: Jossey-Bass.

Rando, W.L. (2001). Writing teaching assessment questions for precision and reflection. In K.G. Lewis (Ed.). *Techniques and Strategies for Interpreting Student Evaluations.* New Directions for Teaching and Learning, No. 87. San Francisco: Jossey-Bass.

Ruthven-Stuart, R. (2004, November). Class Evaluations; how can they improve education? Paper presented at the annual meeting of the Japan Association of Language Teachers (JALT), Nara.

Selden, P. (1993). When students rate professors. *Chronicle of Higher Education,* July 21.

Sinclair, L. and. Kunda, Z (1999). Reactions to a black professional: Motivated inhibition and activation of conflicting stereotypes. *Journal of Personality and Social Psychology, 77*(5), 885-904.

Sinclair, L. and Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me. *Personality and Social Psychology Bulletin, 26*(11), 1329-1342.

Student Assessment of Instruction System (SAIS). (2003). Office of Institutional Research and Assessment, University of Tennessee. Online at: http://oira.tennessee.edu/sais/forms.html.

Student Instructional Report II (SIR II). (2004). Educational Testing Service. Online at: http://www.ets.org/hea/sirII/index.html.

Stake, J.E. (1997). Response to Haskell: Academic freedom, tenure, and student evaluations of faculty. *Educational Policy Analysis Archives, 5*. Online at: http://olam.ed.asu.edu/epaa/v5n8.html.

Svinicki, M.D. (2001). Encouraging your students to give feedback. In K.G. Lewis (Ed.). *Techniques and Strategies for Interpreting Student Evaluations*. New Directions for Teaching and Learning, No. 87. San Francisco: Jossey-Bass.

Theall, M. and Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? In M. Theall, P.C. Abrami, and L.A. Mets, (Eds.). *The Student Ratings Debate: Are They Valid? How Can We Best Use Them?* New Directions for Teaching and Learning, No. 109, San Francisco: Jossey-Bass.

Trout, P.A. (2000). Flunking the test: The dismal record of student evaluations. *The Touchstone, Vol. X*, No. 4, Sept./Oct. Online at: http://www.rtis.com/touchstone/sept00/20flunk.htm.

Wilson, R. (1998). New research casts doubt on value of student evaluations of professors. *The Chronicle of Higher Education*, January 16.

Winskowski, C. and Hanna, C.M. (2004). Junior college students' perception of English skills improvement. *Language and Culture*, No. 6:61-78. Center for Language and Culture Education and Research, Iwate Prefectural University.